How could ego-centric location be defined neurally?

Andrew Glennerster* Miles E. Hansard † Andrew W. Fitzgibbon ‡

Journal of Vision, **1**, 6a (2001) `http://journalofvision.org/1/3/6`

Abstract

The computation of ego-centric location is often assumed to involve a chain of co-ordinate transformations (eye-, head-, body-, world-centred representations) but there are no detailed proposals about how this might be carried out in the visual system (i.e. a shift in the origin and a rotation of the basis vectors describing an entire scene).

Here we present an alternative, which avoids co-ordinate transformations of this type. The model relies on the recognition of lists of objects and the angles required to make saccades from one to the other. The list of currently visible objects surrounding the observer can be split into distant and progressively closer lists of objects. Thus, location is defined hierarchically. As the observer translates, the nearer lists change more rapidly than the distant ones. If the goal is to move between locations defined in this way, we show how retinal flow could have a straight-forward interpretation and that it is positively advantageous for observers to maintain gaze on an object as they translate. We also show how the relationship between locations is affected by the presence of occluding surfaces (such as walls) and how the hierarchy of stored locations could be both progressively extended (to cover wider areas) and refined (to sample space more densely). Eye position signals are not used either in generating or reading out from the representation.

The proposed representation is a set of sensory states linked by motor outputs. This is, at least, something we know the visuo-motor system can store.

Introduction

A question like 'Where am I?' is often assumed to:

- be the responsibility of the dorsal ('where') pathway and the hippocampus (Andersen et al., 1997; Colby, 1998; Burgess et al., 1999) and
- give rise to an answer in a 3-D, world-based coordinate frame.

Here we suggest alternatives to both these assumptions. The goal we set out is to identify which of a set of stored locations corresponds most closely to the observer's current location. We make the following (uncontroversial) points:

*University Laboratory of Physiology, Parks Road, Oxford, OX1 3PT; ag@physiol.ox.ac.uk
†Department of Computer Science, University College London, Gower St, London WC1E 6BT
‡Department of Engineering Science, University of Oxford, Parks Road, OX1 3PJ

- recognition of a list of objects is often sufficient to identify the observer's location (panel 1)
- when this is not the case, recovery of the visual angle between objects may be sufficient (panel 2)
- the precision with which locations of the observer can be distinguished in this way depends on the layout of the points in the scene (panel 3)

The result is a hierarchical system for defining location. More distant points in the scene provide the coarsest level of description, with more detail provided by the nearer points. Stored (i.e recognisable) locations are not 3-D coordinates in this scheme. Instead they are lists of objects and the visual angles between then. The lists have a natural hierarchical structure and could be used to guide navigation between the stored locations. View graph methods of navigation are based on this principle (e.g. Koenderink and van Doorn, 1979; Gillner and Mallot, 1998; Arbib, 1999; Franz and Mallot, 2000).

Translation between stored locations requires a change in the visual angles between objects. Maintaining fixation on a point as the observer translates could simplify the monitoring of such changes in visual angle (Glennerster et al., 2001).

None of the above claims are especially surprising. Nevertheless, they differ from the process of photogrammetry whose goal is a description of the 3-D layout of points in a scene and of the observer's location within it (reviewed by Hartley and Zisserman, 2000). A process rather like photogrammetry is often assumed to underlie human visual processing of 3-D structure and location but there is as yet little evidence to support this idea. We argue here that it may be unnecessary.

Panel 1

Question: 'Where am I?'
Possible answers: $\{loc_1, loc_2, loc_3, loc_4\}$

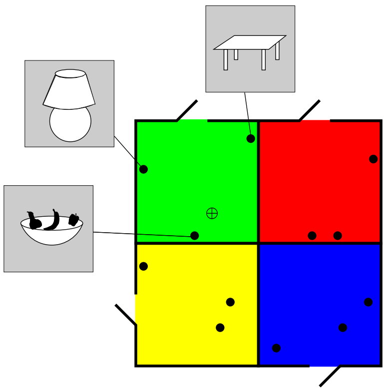

Figure 1: Plan view of four rooms. Here, they are the candidate locations. The lamp, table and fruit bowl uniquely identify the green room

True location: $\oplus$

Stored:

loc_1 : {'lamp', 'table', 'fruit bowl'}
loc_2 : {'lamp', 'desk', 'picture'}
loc_3 : {'table', 'basin', 'fruit bowl'}
loc_4 : {'umbrella', 'table', 'hat'}

Data: {'lamp', 'table', 'fruit bowl'}
Answer: {loc_1}

Details of simulations

In the simulations shown here, we restrict our attention to a two-dimensional world. In this model, scene-features at positions (x_p, y_p), $p = 1, \ldots, P$, are imaged from viewpoint (x, y), and represented by the visual angles

$$\boldsymbol{\rho}_{p,q}(x,y) = \big(\rho_{1,q}(x,y), \ldots, \rho_{p,q}(x,y), \ldots, \rho_{P,q}(x,y)\big), \quad q = 1, \ldots, P,$$

where $\rho_{p,q}$ is the angle between points (x_p, y_p) and (x_q, y_q). In order to avoid the use of a distinguished reference point, we compute all P^2 possible angles from each viewpoint.

As well as general views $\boldsymbol{\rho}(x,y)$, we have R different *reference views*,

$$\big\{\boldsymbol{\rho}_{p,q}(x_1,y_1), \ldots, \boldsymbol{\rho}_{p,q}(x_r,y_r), \ldots, \boldsymbol{\rho}_{p,q}(x_R,y_R)\big\}, \quad q = 1, \ldots, P,$$

taken from distinct locations (x_r, y_r). Each reference view $\boldsymbol{\rho}(x_r, y_r)$ is accompanied by a corresponding list of variances, $\boldsymbol{\sigma}^2(x_r, y_r)$;

$$\begin{aligned}
\boldsymbol{\rho}_{p,q}(x_r,y_r) &= \big(\rho_{1,q}(x_r,y_r), \ldots, \rho_{p,q}(x_r,y_r), \ldots, \rho_{P,q}(x_r,y_r)\big), \quad q = 1, \ldots, P,\\
\boldsymbol{\sigma}^2_{p,q}(x_r,y_r) &= \big(\sigma^2_{1,q}(x_r,y_r), \ldots, \sigma^2_{p,q}(x_r,y_r), \ldots, \sigma^2_{P,q}(x_r,y_r)\big), \quad q = 1, \ldots, P.
\end{aligned}$$

In our simulations, we take $\sigma^2_{p,q} = 1$ in all cases. The fit of the r^{th} reference view to the visual angles obtained at observer-position (x, y) is defined as the squared-difference between $\boldsymbol{\rho}(x,y)$ and $\boldsymbol{\rho}(x_r, y_r)$, summed over all angles,

$$E_r(x,y) = \sum_{q=1}^{P} \sum_{p=1}^{P} \frac{1}{\sigma^2_{p,q}(x_r,y_r)} \big(\rho_{p,q}(x,y) - \rho_{p,q}(x_r,y_r)\big)^2.$$

We use E_r to compute the likelihood of the current view $\boldsymbol{\rho}(x,y)$, under the hypothesis that the viewpoint coincides with that of model view $\boldsymbol{\rho}(x_r, y_r)$. Specifically, we represent the likelihood as $\mathsf{L}\big(\boldsymbol{\rho}(x,y) \mid \boldsymbol{\rho}(x_r,y_r)\big) = e^{-E_r(x,y)}$, which is proportional to the *probability* of the r^{th} location-hypothesis;

$$\mathsf{P}\big(x = x_r, y = y_r \mid \boldsymbol{\rho}(x,y)\big) \propto e^{-E_r(x,y)}.$$

The normalizing constant is obtained by integrating P over all viewpoints, (x, y). The figures plot the maximum likelihood at each point (x, y), colour-coded by the reference location r for which $\mathsf{L}\big(\boldsymbol{\rho}(x,y) \mid \boldsymbol{\rho}(x_r,y_r)\big)$ is maximum.

Panel 2

Question: 'Where am I?'
Possible answers: $\{loc_1, loc_2\}$

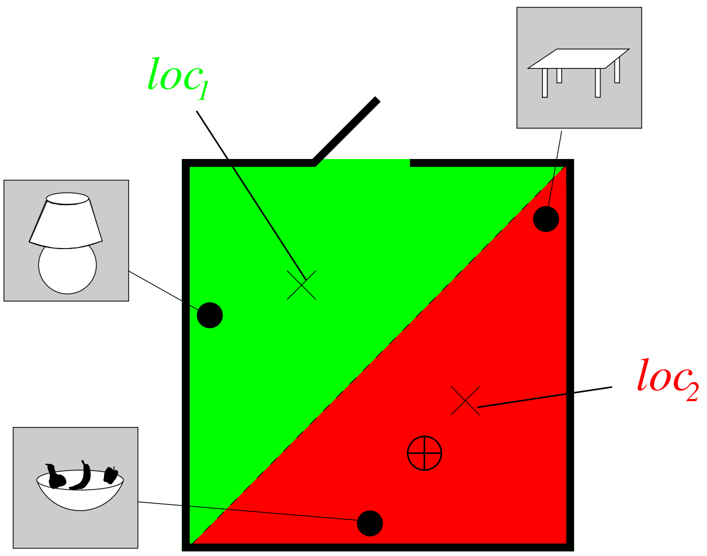

Figure 2: Plan view of one room. Now, two locations on either side of the room must be distinguished. In this case the angles between the lamp, table and fruit bowl are required to make the distinction, as shown below.
True location: $\oplus$

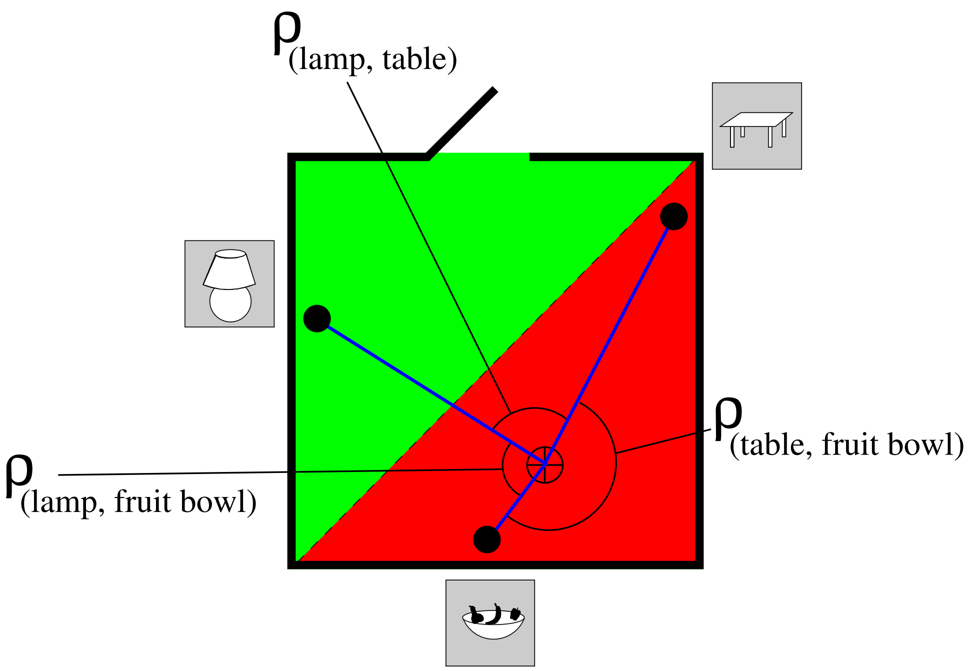

Figure 3: Angles (ρ) between each of the three objects are shown, as viewed from $\oplus$. These are additional data that help distinguish the current location as loc_1 or loc_2. The red and green regions show (very schematically) how the room could be sub-divided according to whether the most likely location is loc_1 or loc_2. A simulation of a similar scene is shown below.

Data: {‘lamp’, ‘table’, ‘fruit bowl’,
ρ(‘lamp’,‘table’),
ρ(‘lamp’,‘fruit bowl’),
ρ(‘fruit bowl’,‘table’)}

Answer: $\{loc_2\}$

<u>A simulation of the above:</u>

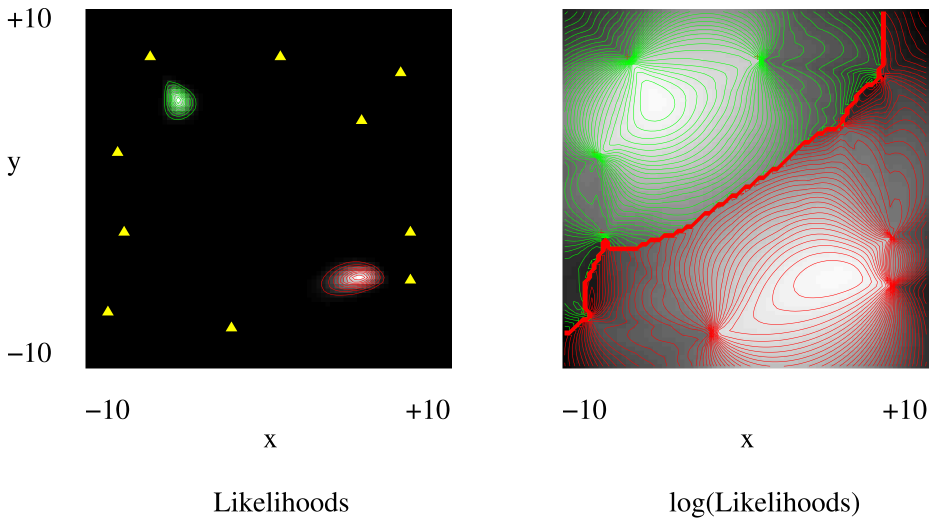

Figure 4: Scene points are shown as yellow triangles. The likelihoods that the observer is at the two candidate locations (loc_1 or loc_2) are plotted in the left-hand plot (see Details panel). On the right, the data are re-plotted showing the log of the likelihoods. This shows the regions of the room for which each of the two reference locations is the most likely to be the observer's current location.

Panel 3

Question: ‘Where am I?’
Possible answers:
$\{loc_1, loc_2, loc_3, loc_4, loc_5, loc_6,$ ‘not sure’$\}$

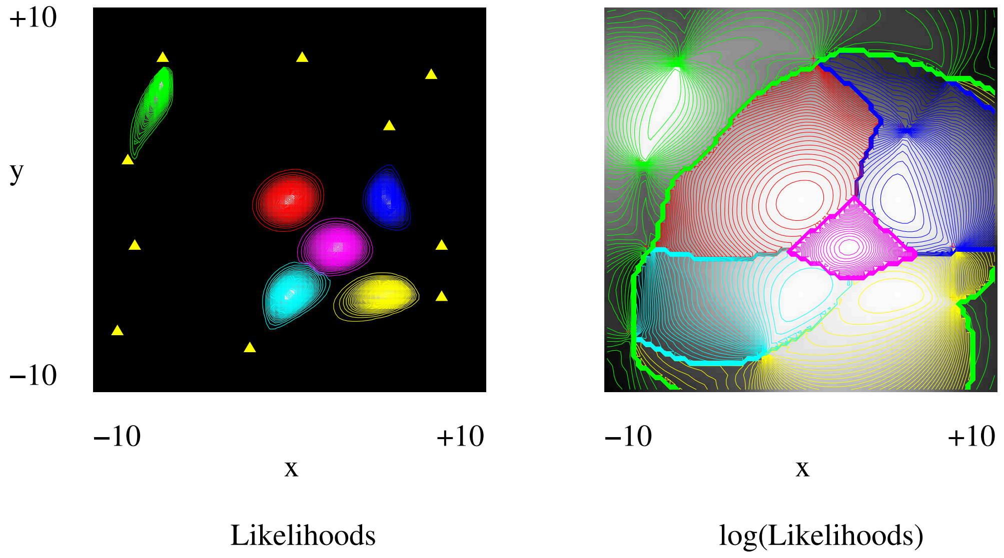

Figure 5: As for figure 4, but now with 6 candidate locations, each identified by a different colour. The scene points are relative near.
Near points constrain the estimate of location more tightly than distant ones, and hence distort the maximum likelihood ‘fields’.
More distant points provide less accurate information about location:

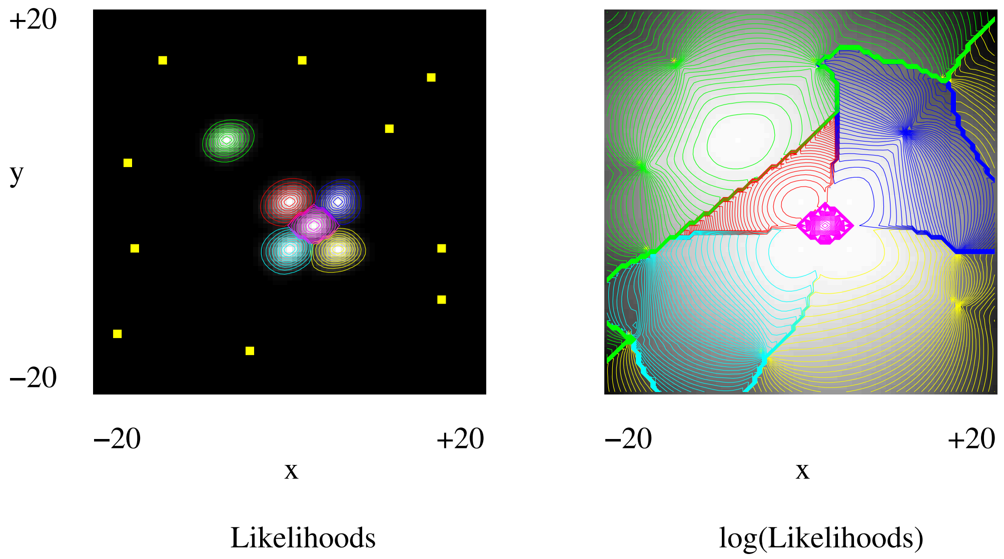

Figure 6: The room is now twice as large and the scene points more distant. The likelihood functions are less distinct.
Even more distant scene points. Locations close to one another are now hard to distinguish:

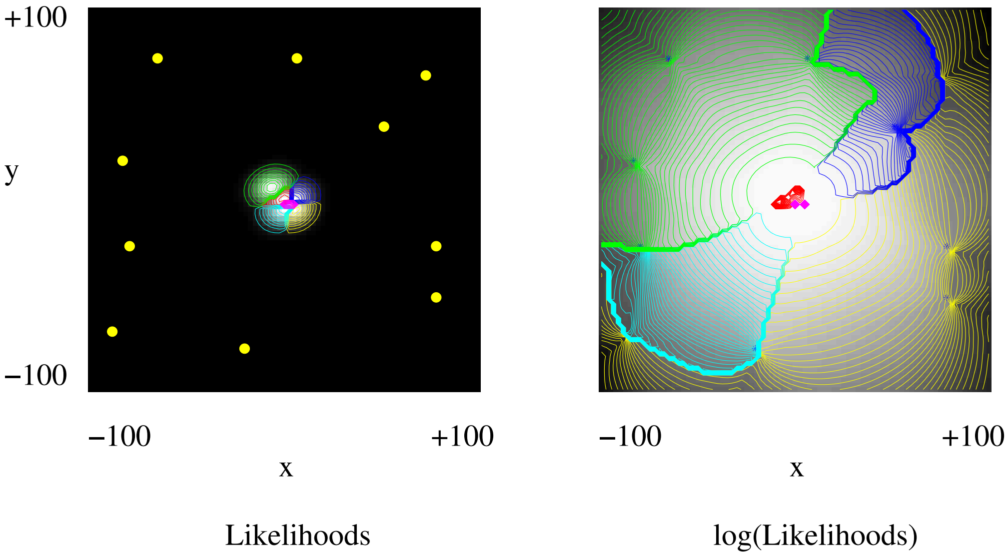

Figure 7: The dimensions of the room are now ten times larger than for figure 5. The scene points are shown as yellow circles. The 6 candidate locations are the same as in figures 5 and 6 but now, because the visual angles between distant objects changes very little as the observer translates, the 6 hypotheses are almost equally likely.

Fixation

Fixating a point (e.g. the lamp) as the observer translates provides a simple way to monitor changes in angles (ρ) with respect to the lamp (e.g. $\rho_{(\text{'lamp', 'table'})}$). Because the lamp is foveated throughout the movement, the motion of the retinal projection of the table signals the change in this angle: $\Delta\rho_{(\text{'lamp', 'table'})}$. Glennerster et al. (2001) discuss this in greater detail (including 2-D retinal projections).

If locations are defined in terms of the relative angles between objects (as suggested here), then monitoring changes in some of these angles (in particular, those measured with respect to the fixation point) could be a useful way to control observer translation between locations.

Conclusion

The scheme for defining ego-centric location outlined here relies on output from the **ventral ('what') stream** of visual processing (panel 1). A list of foveated objects and the angles (saccades) required to foveate each are sufficient to estimate location in the way we have described (panel 2).

In this scheme, the dorsal stream is not involved in the generation of ego-centred representations of space nor the computation of the observer's location in a 3-D frame. Translation of the optic centre (the observer's eye) results in retinal flow, i.e. changes in visual angles between objects. Conversely, since stored locations are *defined* in terms of visual angles between objects, changing the angle between objects is a good way to translate between two locations. Signals from the **dorsal ('where') stream** would clearly be useful in monitoring observer translation in this way.

References

Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J. (1997). Multi-modal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, 20:303–330.

Arbib, M. (1999). Parietal cortex and hippocampus: from visual affordances to the world graph. In Burgess, N., Jeffery, K. J., and O'Keefe, J., editors, *The hippocampal and parietal foundations of spatial cognition*, pages 416–442. Oxford: OUP.

Burgess, N., Jeffery, K. J., and O'Keefe, J. (1999). *The hippocampal and parietal foundations of spatial cognition*. Oxford: OUP.

Colby, C. L. (1998). Action-oriented spatial reference frames in cortex. *Neuron*, 20:15–24.

Franz, M. O. and Mallot, H. A. (2000). Biomimetic robot navigation. *Robotics and Autonomous Systems*, 30:133–153.

Gillner, S. and Mallot, H. A. (1998). Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of Cognitive Neuroscience*, 10:445–463.

Glennerster, A., Hansard, M. E., and Fitzgibbon, A. W. (2001). Fixation could simplify, not complicate, the interpretation of retinal flow. *Vision Research*, 41:815–834.

Hartley, R. and Zisserman, A. (2000). *Mulitple view geometry in computer vision*. Cambridge, UK: Cambridge University Press.

Koenderink, J. J. and van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216.

Acknowledgements

Funded by the Royal Society, UK.